Lexical Emphasis Detection in Spoken French Using F-BANKs and Neural Networks

Abdelwahab Heba[1,2(✉)], Thomas Pellegrini[2], Tom Jorquera[1], Régine André-Obrecht[2], and Jean-Pierre Lorré[1]

[1] Linagora, Toulouse, France
{aheba,tjorquera,jplorre}@linagora.com
[2] IRIT, Université de Toulouse, Toulouse, France
{aheba,pellegri,obrecht}@irit.fr

Abstract. Expressiveness and non-verbal information in speech are active research topics in speech processing. In this work, we are interested in detecting emphasis at word-level as a mean to identify what are the focus words in a given utterance. We compare several machine learning techniques (Linear Discriminant Analysis, Support Vector Machines, Neural Networks) for this task carried out on SIWIS, a French speech synthesis database. Our approach consists first in aligning the spoken words to the speech signal and second to feed classifier with filter bank coefficients in order to take a binary decision at word-level: neutral/emphasized. Evaluation results show that a three-layer neural network performed best with a 93% accuracy.

Keywords: Emphasized content recognition · Non verbal information in speech · SIWIS French speech synthesis database

1 Introduction

Speech in human communication is not only about the explicit message conveyed or the meaning of words, but also includes information, intentionally or not, which are expressed through nonverbal behaviors. Verbal and non-verbal information shape our interactions with others [4]. In [25], for instance, an appropriate use of emphasis was shown to improve the overall perception of synthesized speech. Word-level *emphasis* is considered as an important form of expressiveness in the speech synthesis field with the objective of drawing the listener attention on specific pieces of information.

A speech utterance may convey different meanings according to intonation. Such ambiguities can be clarified by emphasizing some words in different positions in a given utterance. Automatically detecting emphasized content may be useful in spoken language processing: localizing emphasized words may help speech understanding modules, in particular in semantic focus identification [15].

In speech production, various processes occur at word, sentence, or larger chunk levels. According to [6], the tonal variation, defined by pitch variation,

DOI: 10.1007/978-3-319-68456-7 20

is considered as a type of pronunciation variation at the suprasegmental level. Generally, systems for automatic classification of accented words use prosody, and typically use combination of suprasegmental features such as duration, pitch, and intensity features [5,14,17,20,24]. Emphasis cues found in natural speech are more vague and heavily affected by suprasegmental features. Compared to the intensity, pitch and duration are more insensitive to the channel effects such as the distance between the speaker and the microphone. Furthermore, rather than intensity, changes in vocal loudness also affect features such as spectral balance, spectral emphasis or spectral tilt, which were explored in the detection of prominent words [3], focal accent [9], stressed and unstressed syllables [18,19,22]. Indeed, these measures were also generally found to be more reliable than intensity.

In this paper, a statistical approach that models and detects word-level emphasis patterns is investigated. Related works are dedicated to the detection of lexical stress and pitch accent detection, in particular for Computer-Assisted Language Learning [12,13,21,27,28]. The present study differs from these works from the fact that we target at detecting acoustic emphasis at lexical level and in native speech. We plan to detect emphasis at word-level as a first step for future applications we would like to address. In particular, we would like to study if keyword detection in speech transcripts could be improved using a measure of emphasis as an additional piece of information.

Our methods consists first in aligning the speech signal to the spoken words, second in classifying each word segment as emphasized or neutral using filter-bank coefficients (F-BANKs) as input to a classifier. These acoustic features measure the energy from a number of frequency bands and take time dynamics into account. Furthermore, our preliminary experiments showed that F-BANKs outperform the use of single pitch variations (F0). We compare several types of classifiers for this task. As will be reported in this paper, neural networks performed the best.

The present article is structured as follows. Section 2 describes our methodology for word-level emphasis detection, including feature extraction and model description. In Sect. 3, we present the SIWIS French speech synthesis database, then we report a comparison of approaches and analyze the classification results.

2 Method

Figure 1 illustrates the global system schema for an example sentence: "ce FICHIER facilitera principalement la recherche [...]" (*"this FILE will mainly ease the search for [...]"*). In this sentence, the word "FICHIER" (*FILE*) was emphasized by the speaker. As a first step, a word alignment is carried out, which automatically aligns the expected text to the audio speech signal. Then, low-level acoustic features described hereafter are extracted and fed to a binary classifier that takes decisions on the emphasized/neutral pronunciations at word-level.

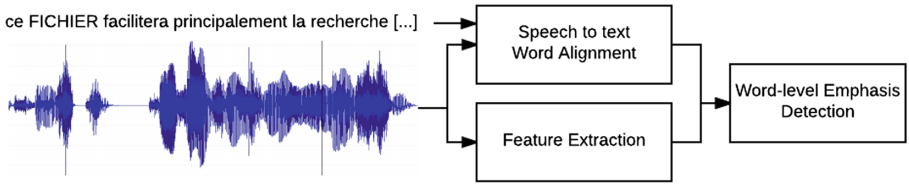

Fig. 1. Word-level emphasis detection

2.1 Word Alignment

We adopt a standard approach used in speech recognition called the time alignment procedure. This procedure is accomplished using supervised phone-based recognition and produces phone-by-phone time markings, which are reduced to a word-by-word format involving the following steps [23]:

- create a word-level grammar from the orthographic transcription (read speech);
- extract acoustic features from the speech signal;
- associate a phone transcription to each word, either extracting it from our pronunciation lexicon or generating them automatically with a grapheme-to-phoneme model; several pronunciations may be associated to a given word;
- perform the word alignment;
- extract the time markings from the aligned word segments.

We used in-house acoustical models trained with the Kaldi Speech Recognition Toolkit [16]. They were trained with the ESTER corpus [8] which consists of 90 h of French broadcast news speech, each broadcast session contains from 20 to 40 min of spontaneous speech. Non-speech sounds, such as breath noises and laughter are indicated in the transcriptions and we explicitly modeled them.

We followed a standard Kaldi recipe to train the models to obtain triphone Gaussian Mixture Models/Hidden Markov Models, on 39 static, delta, and delta-delta Mel-frequency cepstral coefficients, with LDA-MLLT and Speaker Adaptive Training (SAT). Finally, we obtain triphone with about 150 k Gaussian mixtures and 21.2 k HMM states.

Regarding the pronunciation lexicon, we used the 105 k entry CMU-Sphinx French dictionary. For out-of-vocabulary words, pronunciations were derived from a grapheme-to-phoneme tool trained over the CMU-Sphinx lexicon [2]. This concerned a set of 471 words over the 33,628 different word types contained in the SIWIS corpus used in this work.

2.2 Features

As input to the emphasis/neutral classifiers, we use 26 log filter-bank coefficients (F-BANKs) extracted on 25 ms duration frames with a hop size of 10 ms. Different numbers of filter bands were tested, but the set of 26 static F-BANK features has shown to perform well.

For each word of duration N frames, a $N \times 26$ matrix is extracted. We compare the use of these length varying input matrices and the same matrices but in which some context is added: we add left and right frames to reach a $1s$ total duration, which gives a 108×26 matrix for each word.

These matrices are used in two ways, as an image (dimension: 108×26) fed to a Convolutional Neural Network (CNN), or as global statistics features (one value per filter-bank coefficient along time): the minimum, maximum, mean, median, standard deviation, skewness and kurtosis (dimension: 7×26), which are expected to characterize the behavior of each F-BANK coefficient to improve the temporal modeling.

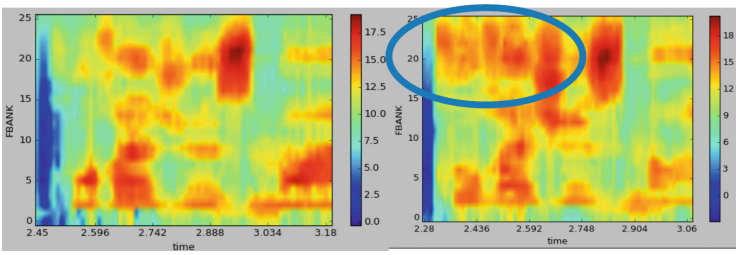

Fig. 2. The two figures represent the F-BANK coefficients for the word: "*courageuse-ment*" (*courageously*) with and without focus emphasis on the right and on the left, respectively. (Color figure online)

Figure 2 shows the 26 F-BANK coefficient images for both an emphasized and a neutral pronunciations of the word "courageusement" (*courageously*). We can notice that the high frequency region on the right figure (with emphasis) has higher energy values, particularly at the beginning of the word, as depicted with a blue ellipse over this area [7].

2.3 Models

Two types of neural networks were tested for our task: a neural network with fully-connected layers (FCNN), and a convolutional neural network (CNN).

In the case of FCNNs, the input layer is the concatenation of the global statistics on the 26 F-BANKs, i.e. a vector of size $7 \times 26 = 182$. The k_0 Softmax outputs estimate the emphasis of each trial. We use rectified linear (ReLU) units that have shown accurate performance in speech recognition tasks [26]. Furthermore, we experimented different number of hidden layers and different number of units. We report results with a single layer and three hidden layers each comprised of 200 units.

With CNNs, the input layer is composed of 108 frames of 26 log filter bank coefficient. Three convolution layers were respectively applied: the frequency filtering 1×26, then dynamic time filtering 108×1 and, finally, 3×3 squared filters. Followed by 2×2 down-sampling (max-pooling) layers, and produce respectively

32, 16, and 8 activation maps that serve as input parameters for three 200-unit dense hidden layers with rectified linear unit (ReLU) activation function. Finally, the output dense layer comprises 2 units with a Softmax activation function to provide a probability.

The networks were trained with the Adam optimization [11] using a cross-entropy cost function. The regularization L^2 was used over all hidden layers. Those models are not very deep but appear to be sufficient to get insights on emphasis detection on a small database such as SIWIS. To carry out our work, Tensorflow was used to perform the experiments on a GPU TITAN 1080 device [1].

Support Vector Machines (SVM) with a Gaussian kernel and Linear Discriminant Analysis (LDA) were also used as a baseline of our experiments.

3 Experiments

3.1 Speech Material

The SIWIS French Speech Synthesis corpus contains read speech recorded from a single native female French speaker, who reads texts selected from three different written sources: books from French novels, parliament speeches and semantically unpredictable sentences. These three written sources were divided to six subsets and serve different purposes. In our study, we only use the sentences containing emphasized words (named "part 5") and their corresponding neutral sentences (contained in parts 1 to 4). The corpus contains $1575 * 2$ sentences equivalent to 3 h 35 min duration of audio, moreover, emphasized words can be seen at different positions in the sentences (begin, middle and end). The manual annotations of emphasized phones are available in the HTS label format. Indeed, SIWIS aims at building TTS systems, investigate multiple styles, and emphasis. For more information about the corpus, the reader may refer to [10].

Word Alignment Experiments. Since the manual annotation provided with the SIWIS database did not allow to get time markings at word-level easily, one needed to perform word alignment as a pre-processing step for emphasis detection. We have grouped the manual time markings of emphasized phones for each word to evaluate the root mean square difference between the manual and the automatic word boundaries, according to the following formula, in which t_m^i and t_a^i are the manual and automatic time markers for the i^{th} word, respectively, and N the total number of words to be aligned:

$$\mathrm{RMS} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(t_m^i - t_a^i)^2}$$

We worked on $1817 * 2$ words (emphasized and neutral pronunciations). The mean duration of these words is 0.372 s ($\pm$19% std), the word "*monoparentales*" (*uniparental*) has the longest duration of about 1 s, and the shortest word "*un*" (*a*) has a 0.03 s duration.

The RMS obtained was 0.243. The smaller, the better the alignments. Nevertheless, we are aware that this value *per se* is difficult to interpret without any other reference value obtained on other speech data. Furthermore, the phone error rate is 8.7%.

A set of 1695 emphasized words were found when grouping phones (*e.g.* "temps en temps" was considered as one word by the corpus annotators), but our manual re-checking of the 1575 sentences lead to 1817 emphasized words. This increase is due to the fact that we consider each word in contiguous emphasized word sequences as several emphasized words (we consider *"temps en temps"* as three different words).

3.2 Classification Results

In this part, we evaluate the word-level emphasis detection method, which consists in applying the procedure shown in Fig. 1. The dataset contains $1817 * 2$ words (emphasized and neutral). The data was split into a training and a test subsets in 80%/20% proportions, respectively, and we performed a 5-fold cross-validation. We chose to keep pairs of the same words with emphasized and neutral pronunciations in the same subset, either in a training or a test fold.

In a first experiment, we focused on the global statistical features extracted over the F-BANKs. As explained previously, they were extracted with and without adding context:

- with context: all the feature matrices share the same 108×26 dimension,
- without context: the feature matrices have a variable time length according to each word: $N \times 26$.

With the different machine learning algorithms used for the classification task, we show in Table 1 that using a bit of context leads to better results in accuracy. The FCNN with 3 hidden layers with 200 units in each layer, using the ReLU activation function, obtained the best performance with a 93.4% accuracy. The variations in performance indicated in the table correspond to the variations according to the five folds used for cross-validation.

Table 1. Accuracy comparison between different classifier types.

With context	No	Yes
FCNN (1 layer)	$81.1 \pm 1.0\%$	$89.9 \pm 1.7\%$
FCNN (3 layers)	$81.3 \pm 5.9\%$	$\mathbf{93.4} \pm 3.3\%$
CNN	_	$90.2 \pm 1.8\%$
SVM	$81.5 \pm 1.0\%$	$92.9 \pm 2.3\%$
LDA	$76.8 \pm 2.4\%$	$89.0 \pm 3.6\%$

In a second experiment, we tested the use of a CNN model. As we showed in Table 1, using context allowed better performance on this task. Consequently,

we used the F-BANK images with context as input to a CNN (matrices of shape 108×26, which represent the extracted F-BANKs features over 1 s of speech signal).

In order to train a CNN model, we needed a validation subset so that we used 70% for training, 10% for validation, and 20% for testing and always with 5 folds. During training, we noticed a clear overfitting of the model on the validation subset so that we used L^2 regularization to overcome this issue. The averaged performance on the five folds was 90.2% (±1.8%), which is not as good as the SVM (92.9%) and the FCNN (93.4%).

3.3 Error Analysis

In this subsection, we analyze the errors made by the best classifier, the FCNN.

A first interesting cue concerns the influence of word duration on performance. The mean duration of the wrong predicted words is about 200 ms. The false positives (emphasized word predicted as neutral) predominantly concern short words such as *"moi"* (me), *"un"* (a), *"pas"* (not), *"cela"* (that), and their mean duration is about 140 ms. On the contrary, the false negatives are longest word mostly, such as *"historien"* (historian), *"constamment"* (constantly). Mean duration of the false negatives is around 400 ms.

By listening to some word utterances incorrectly predicted as neutral, we noticed that the relative focus on these words was as obvious as other emphasized realizations. Smaller intensity values can also be observed in their corresponding spectrograms.

We also explored if there were any relation between the word positions in sentences and the detection errors. Figure 3 shows histograms counting the errors (in black the false positives, in orange the false negatives) according to the word position: at the beginning, middle, or end of a sentence. No clear impact of word position can be observed. Nevertheless, it seems that more false negatives (emphasized words predicted as neutral) occur at the beginning and end of sentences.

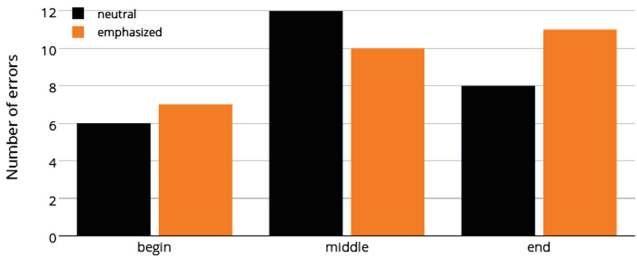

Fig. 3. Number of incorrect predictions according to the word position in sentences. (Color figure online)

4 Conclusions

In this paper, we presented an approach to detect emphasis/neutral intonation at word level specific to the French language. As a first step, a word alignment is carried out, which automatically aligns the expected text to the audio speech signal. Then, F-BANK coefficients are extracted and fed to a binary classifier that takes decisions on the emphasized/neutral decision at word-level.

Evaluation was conducted on SIWIS, a publicly available speech database, that provides read speech material in French with a sub-part manually annotated in terms of emphasis.

Several types of classifiers were tested and the best performance was obtained with a neural network comprised of three fully-connected layers of 200 units each.

As future work, we plan to exploit this system to attempt to improve our keyword extraction module applied to speech transcripts in spoken French. Additionally, we would like to use sequence modelling approach to carry out the detection of emphasized words through entire sentences.

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., et al.: Tensorflow: large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467 (2016)
2. Bisani, M., Ney, H.: Joint-sequence models for grapheme-to-phoneme conversion. Speech Commun. **50**(5), 434–451 (2008)
3. Campbell, N.: Loudness, spectral tilt, and perceived prominence in dialogues. In: Proceedings ICPhS, vol. 95, pp. 676–679 (1995)
4. Campbell, N.: On the use of nonverbal speech sounds in human communication. In: Esposito, A., Faundez-Zanuy, M., Keller, E., Marinaro, M. (eds.) Verbal and Nonverbal Communication Behaviours. LNCS, vol. 4775, pp. 117–128. Springer, Heidelberg (2007). doi:10.1007/978-3-540-76442-7_11
5. Campbell, W.N.: Prosodic encoding of English speech. In: Second International Conference on Spoken Language Processing (1992)
6. Cohn, A.C., Fougeron, C., Huffman, M.K.: The Oxford Handbook of Laboratory Phonology. Oxford University Press, Oxford (2012). Sect. 6.2, pp. 103–114
7. Cole, J., Mo, Y., Hasegawa-Johnson, M.: Signal-based and expectation-based factors in the perception of prosodic prominence. Lab. Phonol. **1**(2), 425–452 (2010)
8. Galliano, S., Geoffrois, E., Mostefa, D., Choukri, K., Bonastre, J.F., Gravier, G.: The ESTER phase II evaluation campaign for the rich transcription of French broadcast news. In: INTERSPEECH, pp. 1149–1152 (2005)
9. Heldner, M.: On the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in swedish. J. Phon. **31**(1), 39–62 (2003)
10. Honnet, P.E., Lazaridis, A., Garner, P.N., Yamagishi, J.: The SIWIS French speech synthesis database? Design and recording of a high quality French database for speech synthesis. Technical report, Idiap (2017)
11. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
12. Li, K., Meng, H.: Automatic lexical stress and pitch accent detection for L2 English speech using multi-distribution deep neural networks. Speech Commun. (2016)

13. Li, K., Zhang, S., Li, M., Lo, W.K., Meng, H.M.: Prominence model for prosodic features in automatic lexical stress and pitch accent detection. In: INTERSPEECH, pp. 2009–2012 (2011)
14. Narupiyakul, L., Keselj, V., Cercone, N., Sirinaovakul, B.: Focus to emphasize tone analysis for prosodic generation. Comput. Math. Appl. **55**(8), 1735–1753 (2008)
15. Noth, E., Batliner, A., Kießling, A., Kompe, R., Niemann, H.: Verbmobil: the use of prosody in the linguistic components of a speech understanding system. IEEE Trans. Speech Audio Process. **8**(5), 519–532 (2000)
16. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al.: The kaldi speech recognition toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, No. EPFL-CONF-192584. IEEE Signal Processing Society (2011)
17. Shriberg, E., Stolcke, A., Hakkani-Tür, D., Tür, G.: Prosody-based automatic segmentation of speech into sentences and topics. Speech Commun. **32**(1), 127–154 (2000)
18. Sluijter, A.M., Shattuck-Hufnagel, S., Stevens, K.N., Van Heuven, V., et al.: Supralaryngeal resonance and glottal pulse shape as correlates of prosodic stress and accent in American English (1995)
19. Sluijter, A.M., Van Heuven, V.J.: Spectral balance as an acoustic correlate of linguistic stress. J. Acoust. Soc. Am. **100**(4), 2471–2485 (1996)
20. Streefkerk, B.M., Pols, L.C., Ten Bosch, L., et al.: Automatic detection of prominence (as defined by listeners' judgements) in read aloud Dutch sentences. In: ICSLP (1998)
21. Tepperman, J., Narayanan, S.: Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners. In: IEEE International Conference on Proceedings of the Acoustics, Speech, and Signal Processing (ICASSP 2005), vol. 1, pp. I–937. IEEE (2005)
22. Van Kuijk, D., Boves, L.: Acoustic characteristics of lexical stress in continuous telephone speech. Speech Commun. **27**(2), 95–111 (1999)
23. Wheatley, B., Doddington, G., Hemphill, C., Godfrey, J., Holliman, E., McDaniel, J., Fisher, D.: Robust automatic time alignment of orthographic transcriptions with unconstrained speech. In: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-1992, vol. 1, pp. 533–536. IEEE (1992)
24. Wightman, C.W., Ostendorf, M.: Automatic labeling of prosodic patterns. IEEE Trans. Speech Audio Process. **2**(4), 469–481 (1994)
25. Yu, K., Mairesse, F., Young, S.: Word-level emphasis modelling in HMM-based speech synthesis. In: 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 4238–4241. IEEE (2010)
26. Zeiler, M.D., Ranzato, M., Monga, R., Mao, M., Yang, K., Le, Q.V., Nguyen, P., Senior, A., Vanhoucke, V., Dean, J., et al.: On rectified linear units for speech processing. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3517–3521. IEEE (2013)
27. Zhao, J., Yuan, H., Liu, J., Xia, S.: Automatic lexical stress detection using acoustic features for computer assisted language learning. In: Proceedings of the APSIPA ASC, pp. 247–251 (2011)
28. Zhu, Y., Liu, J., Liu, R.: Automatic lexical stress detection for English learning. In: Proceedings of the 2003 International Conference on Natural Language Processing and Knowledge Engineering, pp. 728–733. IEEE (2003)